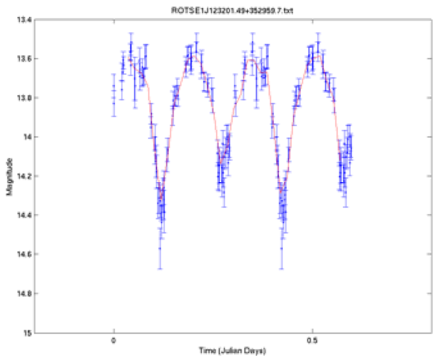
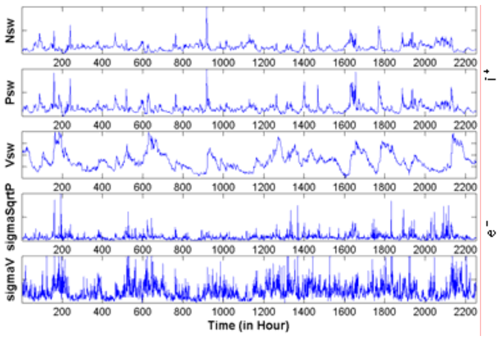
Advanced Machine Learning for Astronomical Time Series Data Analysis

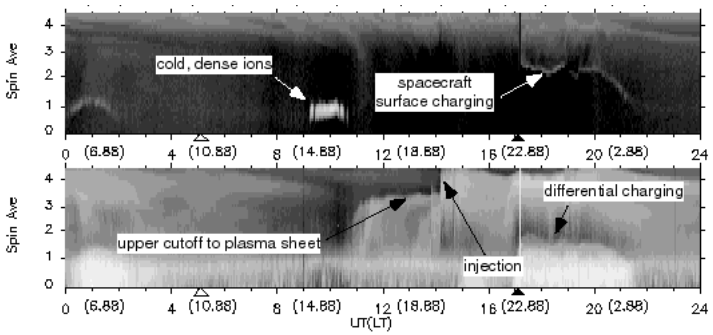
Simon Perkins
Przemek Wozniak
Steven Williams

Los Alamos National Laboratory

Time Series Data Sources





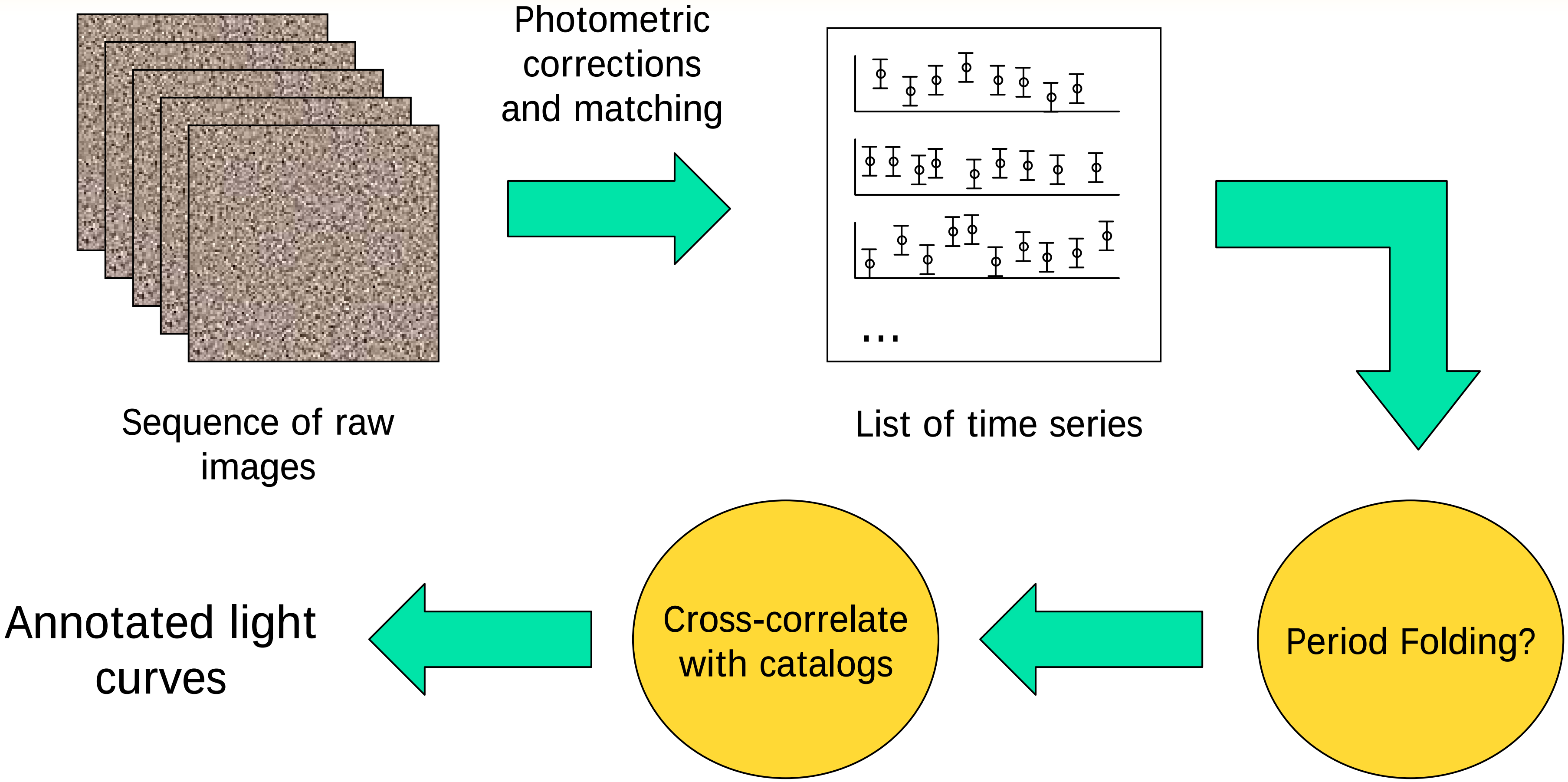


- Our earlier work focused on space physics data.
- More recently we’ve been using astronomical data from sky surveys.
- Cleaner data, better defined problems...

Tasks in Astronomical Time Series Analysis

- Typical data:
  - Repeated observations of a piece of sky over a period of minutes to years.
  - Persistent objects and transient objects.
- Typical tasks:
  - Rejecting “uninteresting” objects.
  - Identifying transient objects.
  - Categorizing persistent objects.
  - Detecting anomalous objects.

Preprocessing Pipeline



Sequence of raw images

Photometric corrections and matching

List of time series

…

Period Folding?

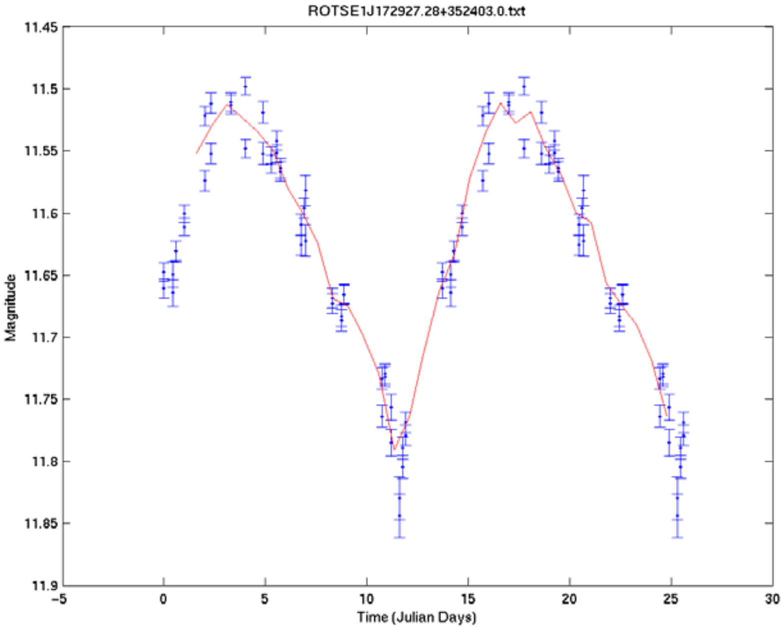
Cross-correlate with catalogs

Annotated light curves

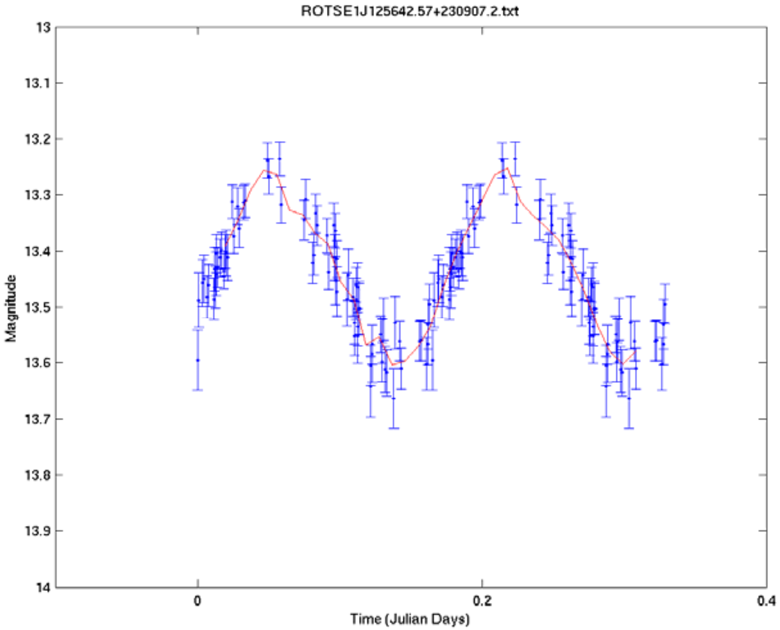
Machine Learning Examples

1. Categorization of ROTSE variable stars.
2. Identification of Miras in ROTSE data.
3. Detection of anomalous Miras

ROTSE Light Curves I

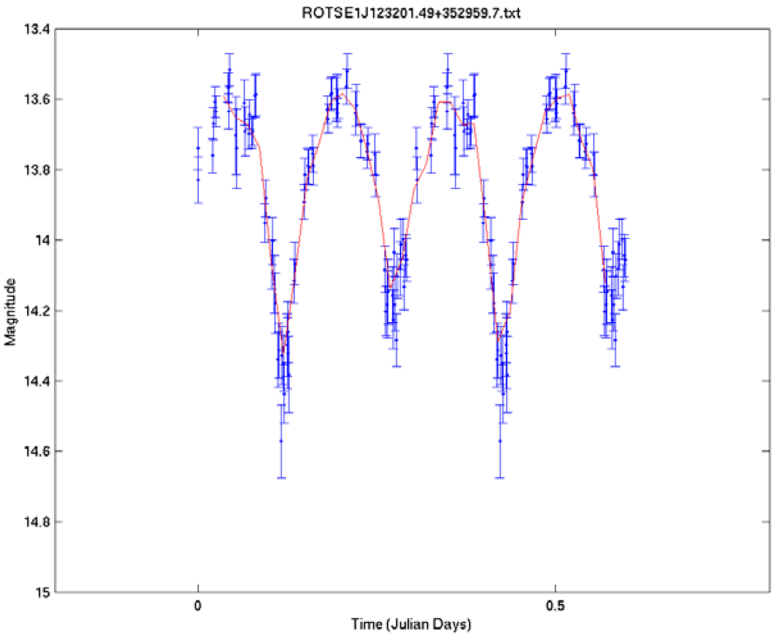


Cepheid

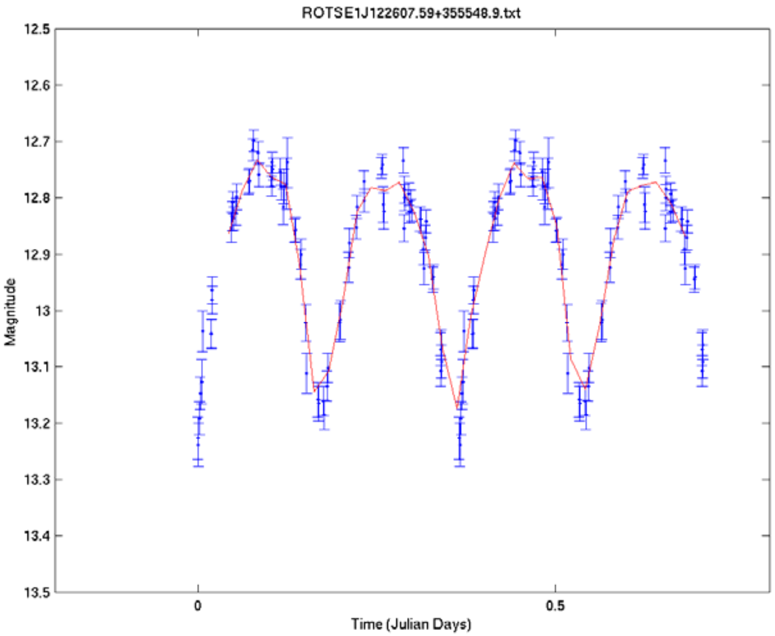


Delta Scuti

ROTSE Light Curves II

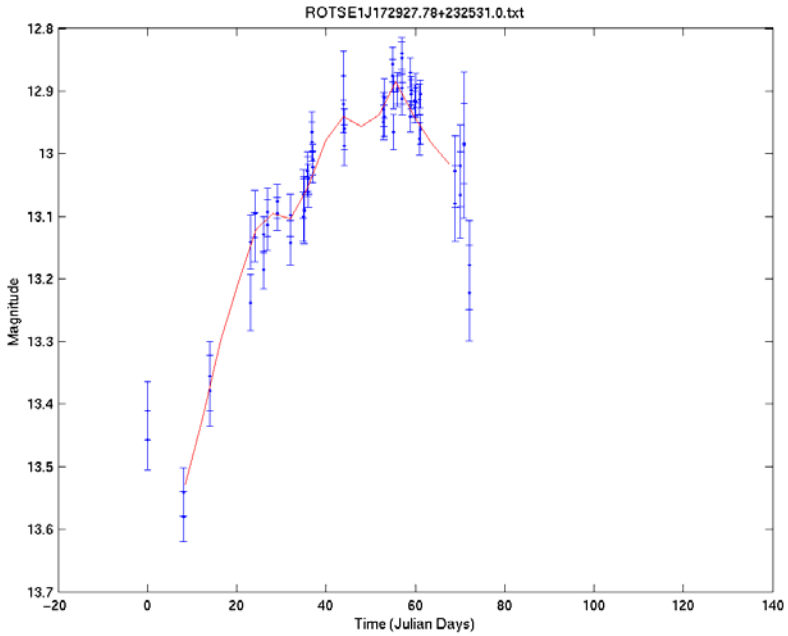


Detached Eclipsing System

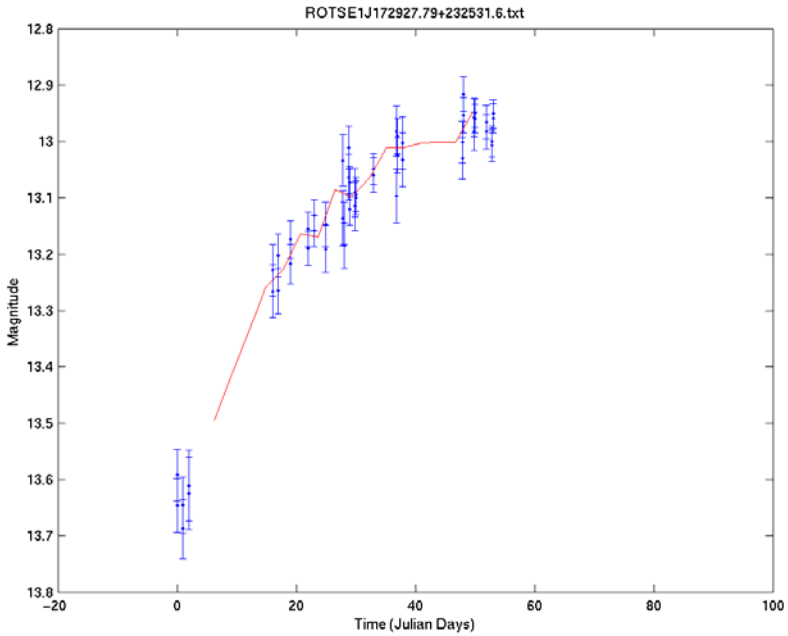


Contact Binary System

ROTSE Light Curves III

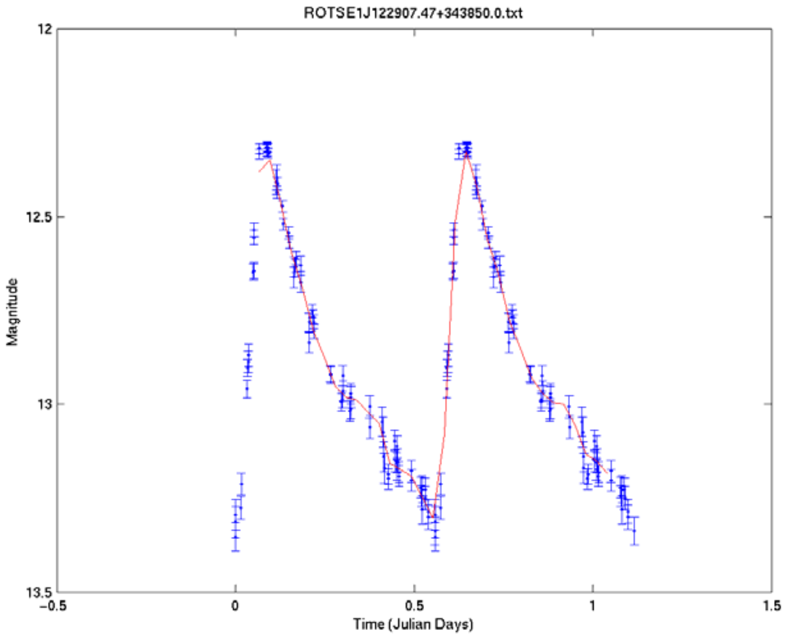


Long Period Variable

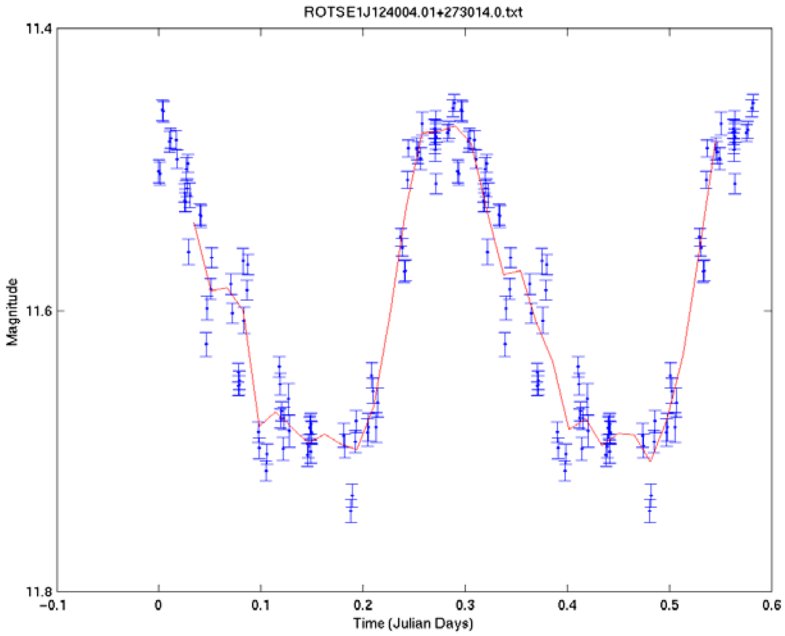


Mira Star

ROTSE Light Curves IV



RR Lyrae Type AB



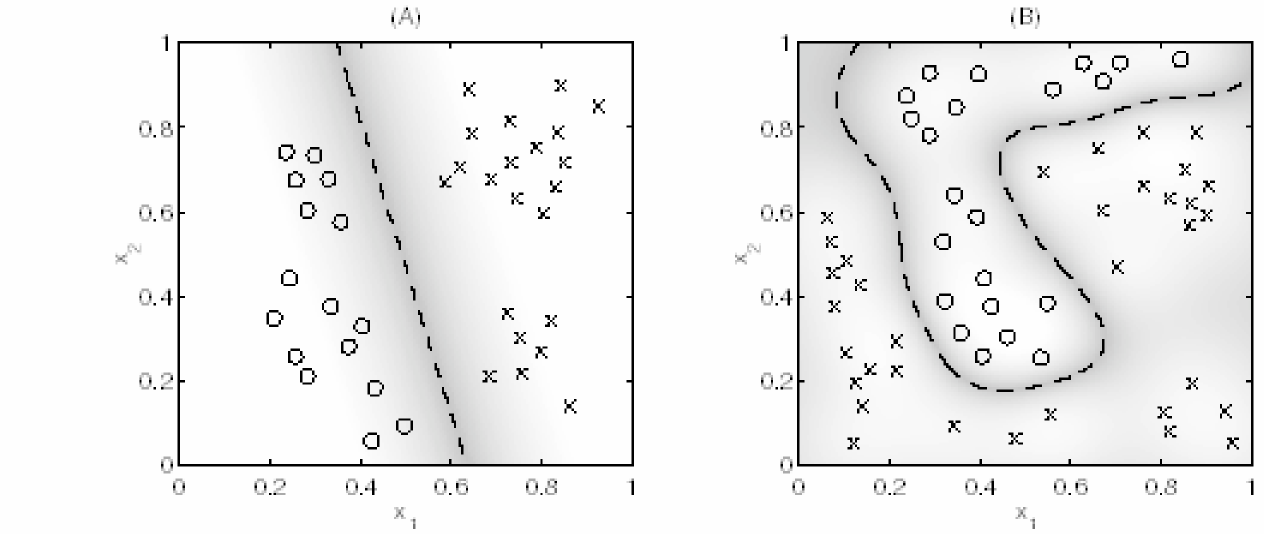
RR Lyrae Type C

Support Vector Machines

- State-of-the-art learning algorithm
- Mathematically well-founded
- Can learn highly non-linear classifiers
- Empirically very successful
- Avoids “overfitting”
- Fast to train

Classification in Feature Space

1. Extract set of numeric features from each star
2. Plot each star as a point in “feature space”
3. Attempt to find a “maximal margin” discriminant separating the classes



General Principle

- How to fit complex high dimensional data without overfitting?
- Combine:
  - Very flexible model.
  - Measure of “capacity” / “complexity”.
- Optimize weighted sum of:
  - Error on training data.
  - Complexity measure of model.

Experiments

- Based on Tim McKay's published collection of almost 2000 variable star light curves.
- Attempt to use SVM for three tasks:
  - Discriminate RRAB stars from all other classes.
  - Discriminate RRAB stars from RRC stars.
  - Discriminate all stars into correct categories.
- Use McKay's published class labels.

Training Set Details

- 1923 variables in total:
  - 209 Cepheids
  - 103 Delta Scuti
  - 127 Detached Eclipsing
  - 419 Contact Binaries
  - 577 Long Period Variables
  - 162 Mira Stars
  - 204 RR Lyrae Type AB
  - 123 RR Lyrae Type C
- Two thirds of data used for training, one third for testing.

ML Details

- Features: period, oscillation amplitude, and magnitude and phase of first eight Fourier components.
- Used LIBSVM – public domain software.
- Gaussian kernel
- C = 10 (found by quick trial and error)
- Three way cross-validation used to get unbiased estimate of prediction accuracy.

Results

- RRAB vs all others: 95.4% accuracy.
- RRAB vs RRC: 93.7% accuracy.
- Full classification into 8 classes: 73.9% (Compare with 12.5% expected randomly)
- Training times of a few seconds.
- Note that accuracy scores are “out of training sample” percentages.

Miras

- Long period red variables.
- Miras, Semi-regulars and Irregulars.
- Interesting role as a “standard candle”.
- Strange flaring events noted in a few Miras.

Identifying Miras

- ROTSE data from Northern Sky Variability Survey.
- 20 million light curves analyzed.
- Manual “cuts”:
    - Reject stars with low variablity -> 98,000
    - Reject stars with rapid variation -> 9,371
    - Correlate with 2MASS to get colors ->8,678
- SVM then used to identify Miras from amplitude, period and color information, using 2500 matching stars in GCVS catalog as training data.
- Approximately 1,100 new Miras found (doubling the number of known Miras).
- Results published in ApJ (Wozniak et al.).

Anomalous Miras

- A few Miraes show occasional small flaring events superimposed on top of regular large variation. Can we identify these?
- We have fitted flexible models to Mira light curves using regularization techniques to reduce overfitting (regularized B-splines).
- Looking for small but statistically significant deviations from the smooth model.
- Work in progress...

Future Work

- Further work on Miras.
- Development of regularized modelling techniques as a general tool for time series analysis.
- Automatic discovery of suitable features for classification of time series.